

De-novo assembly using Illumina Genome Analyzer data

Markus Bauer¹, Ole Schulz-Trieglaff¹, Michael Eberle², Nan Leng³,
David W. Williamson³, Anthony J. Cox¹, Dirk Evers¹

1 Introduction

Recent advances in chemistry and data analysis enabled the generation of reads with more than 100 bp length using the Illumina platform. In this poster, we demonstrate the versatility of these reads for assembly.

Until a couple of years ago the field of sequencing was dominated by traditional Sanger sequencing that provides reads that are typically 400-800 nucleotides long. Sanger sequences were used in the de-novo assembly of large mammalian genomes, such as human or mouse, and mostly followed the *overlap consensus approach*.

In the last years, however, emerging next-gen sequencing platforms as Illumina's *sequencing by synthesis* [1] approach generate high quality sequence data orders of magnitude larger than Sanger sequencing. Illumina reads typically range from 35 to more than 100 nucleotides, requiring a change in the strategies for de-novo assembly. Approaches for de-novo assembly using short reads include velvet [5], Euler-USR [2], and Forge [4]. In our experiments we want to compare the performance of these assemblers.

2 Results

We took paired read sequence data from the *Genome Analyzer II* platform and compared velvet, Euler-USR, and Forge on these data. In a first experiment we created an E.coli data set from short (200 bp) and long insert (≥ 2000 bp) protocol data and compared the performance of the assemblers. We were able to build contigs that are several mega bases long, and we verified the quality of the assembly using amosvalidate and MUMmer.

In our second experiment, we used an E. coli library with a maximal fragment size of 250 nucleotides. We sequenced 150 on each side creating overlapping reads. This allowed us to combine the reads creating sequences that are up to 250 bases long. The alignment routines take the base quality scores into account and perform an error correction at the alignment stage. We compare the assemblies to the ones using only non-overlapping reads.

The last part of this poster deals with a long (2000 bp) insert E. coli library with a maximal fragment size of up to 160 nucleotides. The reads being sequenced are each 100 nucleotides long and thus overlap in most cases. Due to specifics of the sample preparation, the reads might contain DNA from non-adjacent parts of the genome. The challenge is to find these junction sites: To this end, we follow an approach similar to that taken by Pevzner et al. in [3]. We counted the occurrences for each k-mer in the input data set and, after performing an error correction, cluster the positions of k-mers that appear less often than we would expect. Contiguous stretches of rarely occurring k-mers provide candidates for these junction sites. We remove putative junction sites and use the remaining sequence parts in the assembly process. In addition, we verify the correctness of the detected junction sites by aligning the resulting sequences to the reference.

References

- [1] D.R. Bentley et al. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry". *Nature*, 456:53–59, 2008.
- [2] M.J. Chaisson and P.A. Pevzner. "Short read fragment assembly of bacterial genomes". *Genome Research*, 18(2):324–330, 2008.
- [3] P.A. Pevzner, H. Tang, and M.S. Waterman. "An Eulerian path approach to DNA fragment assembly". *Proceedings of the National Academy of Sciences of the United States of America*, 98(17):9748–9753, 2001.
- [4] D. Platt and D.J. Evers. "Forge: A Parallel Genome Assembler Combining Sanger and Next Generation Sequence Data". Manuscript, 2009.
- [5] D.R. Zerbino and E. Birney. "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs". *Genome Research*, 18(5):821–829, 2008.

¹Illumina Cambridge, Chesterford Research Park, Cambridge CB10 1XL, United Kingdom.

²Illumina San Diego, 9885 Towne Centre Drive, CA-92121 San Diego, United States.

³Illumina Hayward, 25861 Industrial Blvd., CA-94545 Hayward, United States.